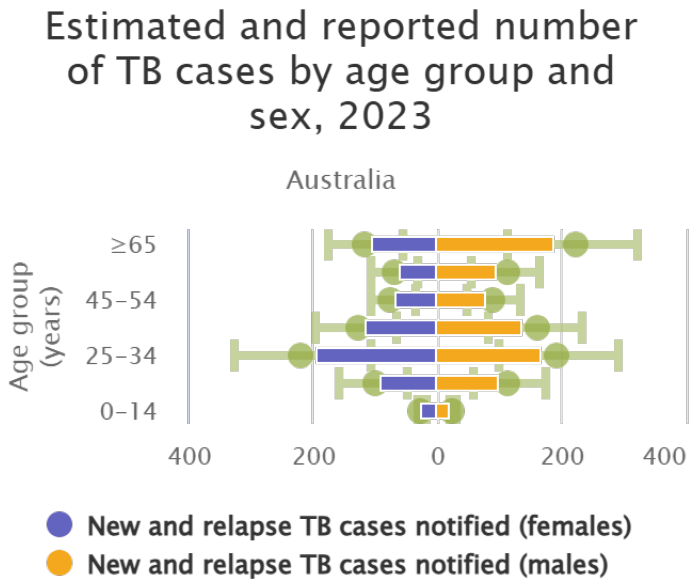
Travis Hornbuckle

Part I – Tuberculosis Project

Introduction to Statistics (MATH1040)

The countries were selected utilizing the Stratified Sampling method, with observational units from WHO member organizations by first grouping them into their assigned WHO regions. Due to the number of countries selected, the South-East Asian and Western Pacific Regions were combined to give equal opportunity for the countries in those regions to be selected. I then utilized GeoGebra's randBetween function to randomly select two numbers that corresponded to a country within those regions when listed alphabetically.

Country	Total TB Incidence	Success Rate	Cohort Size
South Sudan	227	82%	18,865
Botswana	244	76%	3,187
Colombia	46	74%	16,692
Guyana	64	70%	392
Singapore	42	73%	2,596
Australia	6.2	86%	1,187
Azerbaijan	72	82%	3,606
Czech Republic	4.8	63%	360
Egypt	9.2	87%	7,967
Libya	59	60%	2,168



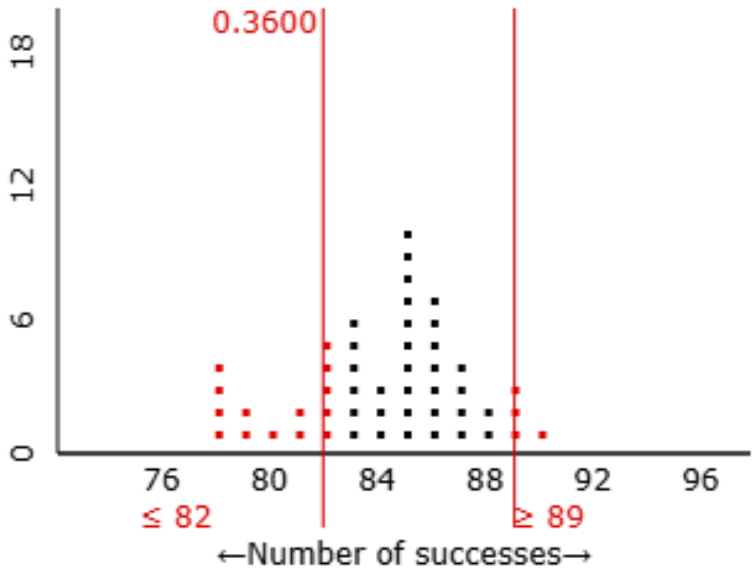
In the chart represented above, we can see that the age group of female Australians with the largest notified cases is 25-34 years old – representing 29.528% (194 cases in the 25-35 age group / 657 female notified cases) of the female notified cases in Australia. On the reverse, we can also see that although Tuberculosis is more commonly reported in men, the smallest number of reported cases in the country comes from Australian men, aged 0-14 years old. This is representing that only 2.445% (19 reported cases / 777 male notified cases) of reported cases are within that age and gender range.

If the sample size of this data set was at 5,000 people total, across the sexes, here is how I would expect the numbers of reported cases to fall based on age, with the highlighted cells representing the highest projected reported cases in females by age group, and the smallest reporting group in males. I calculated this by dividing the number of reports from each age/sex group by the total amount of reported cases – this gave me the proportion of each group to the total number of reported cases. I took this proportion and multiplied it by new population of 5,000 to get the new distribution of cases over each group as seen below.

PROJECTED TB Case Distribution by Age (Cohort of 5,000)

Age group (years)	N&R (Females)	% RPT of Sex	Age/Sex % of Cohort	N&R (Males)	% RPT of Sex2	Age/Sex % of Cohort3
0-14	105	3.805%	2.106%	80	2.445%	1.601%
15-24	383	13.851%	7.666%	413	12.613%	8.256%
25-34	817	29.528%	16.344%	708	21.622%	14.153%
35-44	489	17.656%	9.773%	569	17.375%	11.373%
45-54	278	10.046%	5.560%	320	9.781%	6.403%
55-64	257	9.285%	5.139%	392	11.969%	7.835%
≥65	438	15.830%	8.762%	792	24.196%	15.838%
Total	**2767**			**3273**		

Tuberculosis has been an ongoing issue in the world for quite some time. As we continue to develop as countries, organizations have created ways of ranking countries based on their ability to treat individuals with Tuberculosis. I would like to evaluate **South Sudan's** Tuberculosis rate to see if it is different from the expectation of 85%. To test this, our null hypothesis would be: H_0: $p = 0.85$ – to see if South Sudan's TB Rate is equal to the 85% threshold, and our alternative hypothesis would be H_A: $p \neq 0.85$ to see if South Sudan has a rate that differs from the expectation. I expect that with South Sudan I will reject the null hypothesis that $p = 0.85$ because there will be sufficient evidence to support the alternate hypothesis that $p \neq 0.85$. With this test I will use an alpha level of 0.05 ($\alpha = 0.05$) to set the significance level in a way that avoids errors.



Based on the outcome of the experiment I can see that my p-value – or the probability of the data being favorable to the alternative hypothesis is 0.36 ($p = 0.36$), and the alpha level of 0.05 ($\alpha = 0.05$). With $p > \alpha$, or the probability of the data being favorable to the null hypothesis being higher than the level of significance, we can determine that there is insufficient evidence to

show that the Tuberculosis rate for South Sudan does not equal 85%, and we fail to reject the null hypothesis.

Travis Hornbuckle

Part II – Tuberculosis Project

Introduction to Statistics (MATH1040)

Using the formula ($n \leq 0.5N$), we will test to make sure the observations are independent. For our scenario, $\boldsymbol{n}$ is going to stand in for the cohort size, and the success rate of TB testing in the country will be $\hat{\boldsymbol{p}}$, with our confidence interval at 95%. Here's what that looks like for each of our countries:

EGYPT

$$7{,}967 \leq 0.5N \quad \rightarrow \quad \frac{7967}{0.05} \leq N \quad \rightarrow \quad N \geq 159{,}340$$

$$np \geq 10 \quad \rightarrow \quad (7{,}697) * 0.87 \geq 10 \quad \rightarrow \quad 6{,}696 \geq 10$$

$$n(1-p) \geq 10 \quad \rightarrow \quad 7{,}697(1-0.87) \geq 10 \quad \rightarrow \quad 1{,}001 \geq 10$$

LIBYA

$$2{,}168 \leq 0.5N \rightarrow \frac{2168}{0.05} \leq N \rightarrow N \geq 43{,}360$$

$$np \geq 10 \quad \rightarrow \quad (2{,}168) * 0.6 \geq 10 \quad \rightarrow \quad 1301 \geq 10$$

$$n(1-p) \geq 10 \quad \rightarrow \quad 2{,}168(1-0.6) \geq 10 \quad \rightarrow \quad 867 \geq 10$$

In these three equations, we can see that both countries meet all the criteria to have a valid confidence interval. Assuming both countries have a 95% confidence level, we can now find the upper and lower limits for each country.

EGYPT

Successes	6931
n	7967
SE	0.0038
Lower Limit	0.8626
Upper Limit	0.8773
Interval	0.87 ± 0.0074

LIBYA

Successes	1301
n	2168
SE	0.0105
Lower Limit	0.5795
Upper Limit	0.6207
Interval	0.6001 ± 0.0206

Based on the confidence intervals, we can assume that 85% as a successful treatment threshold is **not** a likely value for either of the countries that I have selected. Based on the excerpts of data above, we can see that both countries are on opposite sides of the

standard treatment threshold, with Libya ranging from 0.579 to 0.62%, and Egypt being above the standard, between 0.862 and 0.877.

Now, previously I looked at South Sudan – how would this information apply to them? Using the hypotheses we already have H_0: $p = 0.85$ and H_A: $p \neq 0.85$, and an **α** of 0.05,

South Sudan

$$18{,}865 \leq 0.5N \rightarrow \frac{18{,}865}{0.05} \leq N \rightarrow N \geq 337{,}300$$

$$np \geq 10 \quad \rightarrow \quad (18{,}865) * 0.82 \geq 10 \quad \rightarrow \quad 15{,}469 \geq 10$$

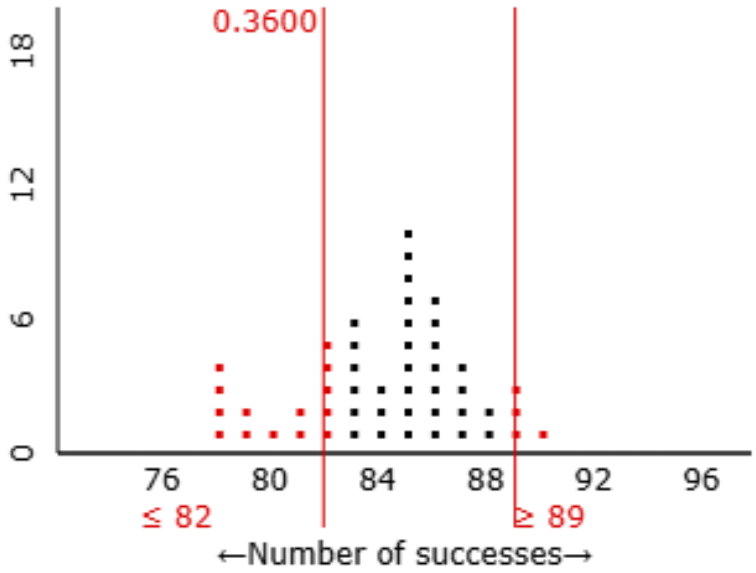
$$n(1-p) \geq 10 \quad \rightarrow \quad 18{,}865(1-0.82) \geq 10 \quad \rightarrow \quad 3{,}396 \geq 10$$

Successes	15469
n	18865
z	-11.5458
p	0

Successes	15469
n	18865
SE	0.0028
Lower Limit	0.8145
Upper Limit	0.8255
Interval	0.82 ± 0.0055

With the outputs above, we can see the p value and the test statistic. Based on the information we have collected, we do not have sufficient evidence to support the alternative, and we fail to reject the null.

Comparing this to my initial hypothesis and testing, I feel that this test is much more accurate and aligns with my previous results. I feel that both tests showed similar data, and though the results are slightly different, both come to the same conclusion.


0.3600
18
12
6
0
76
80
84
88
92
96
≤ 82
≥ 89
←Number of successes→

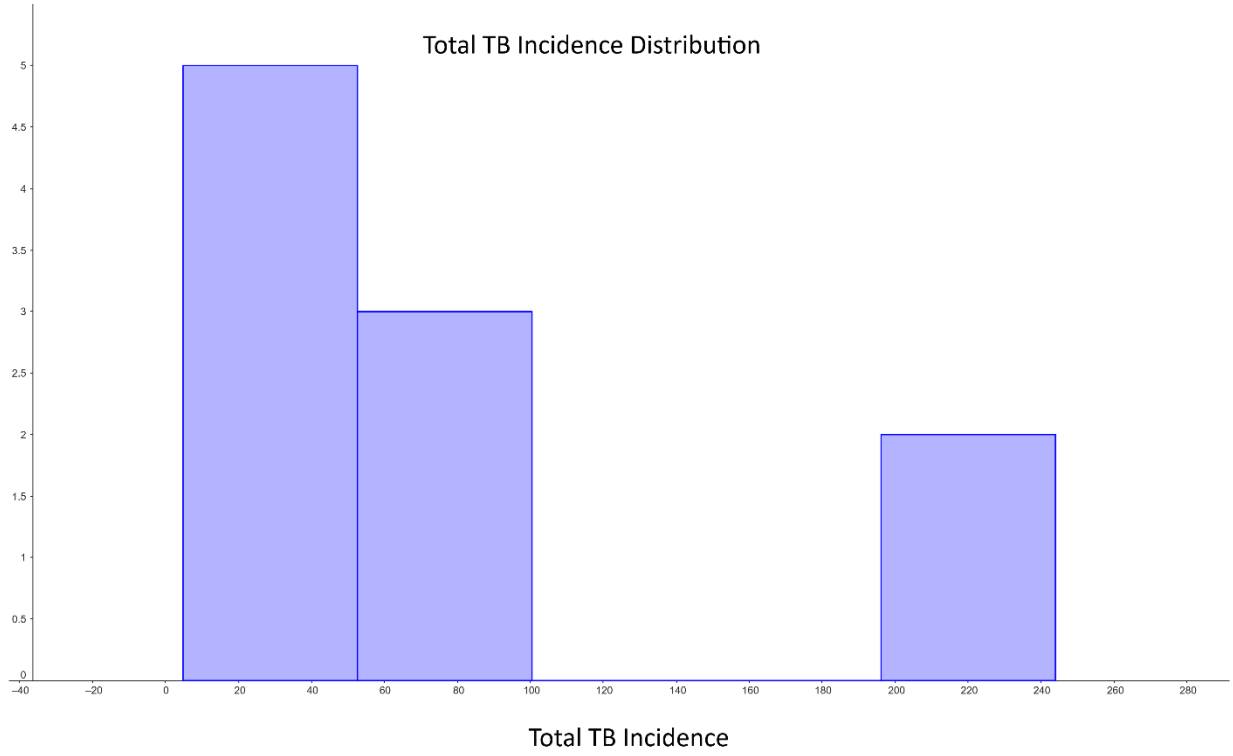
Travis Hornbuckle

Part III – Tuberculosis Project

Introduction to Statistics (MATH1040)

As a reminder, here are the dimensions being evaluated for the selected countries:

Country	Total TB Incidence	Cohort Size
South Sudan	227	18,865
Botswana	244	3,187
Colombia	46	16,692
Guyana	64	392
Singapore	42	2,596
Australia	6.2	1,187
Azerbaijan	72	3,606
Czech Republic	4.8	360
Egypt	9.2	7,967
Libya	59	2,168



When comparing the number of TB Incidence, as represented above, we can see that the distribution is right-skewed, with a mid-point (median) of 52.5 cases and an inter quartile range of 62.8.

With this data, there are two outliers – Both South Sudan (227 Incidences) and Botswana (244 Incidences) are outliers because of their high rate of contraction. We can see this by calculating the interquartile range (IQR = Q_3 - Q_1).

IQR = (72) – (9.2)

IQR = 62.8

We then take the IQR and using the formulas, we can calculate the upper and lower fences: (Lower Fence = Q_1-1.5(IQR)) and (Upper Fence = Q_3+1.5(IQR))

Lower = 9.2 – 1.5(62.8)

Lower = -85

Upper= 72+1.5(62.8)

Upper = 22.6

Now, we would look for the values that are outside of those values. In our case, it is only the two countries that do not fall within the fences.

We also want to compute a confidence interval using the three conditions – we want to check for Random Sampling, Independence, and making sure the populations are normal. To evaluate these conditions – we will use the sum of all the countries populations as n. Therefore, $n = 18865+3187+16692+392+2596+1187+3606+360+7967+2168$, or $n = 57{,}020$.

Random Sampling

We know that the sample is random as the data underwent a random selection process.

Independence

$n_1 \leq 0.05N$ and $\eta_2 \leq 0.05N$

$57{,}020 \leq 0.05N$ and $57{,}020 \leq 0.05N$

$N \geq 1{,}140{,}400$ and $N \geq 1{,}140{,}400$

We know that observations are independent of each other.

Population Size

$n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$

$57{,}020(0.0201) \geq 10$ and $57{,}020(1-0.0201) \geq 10$

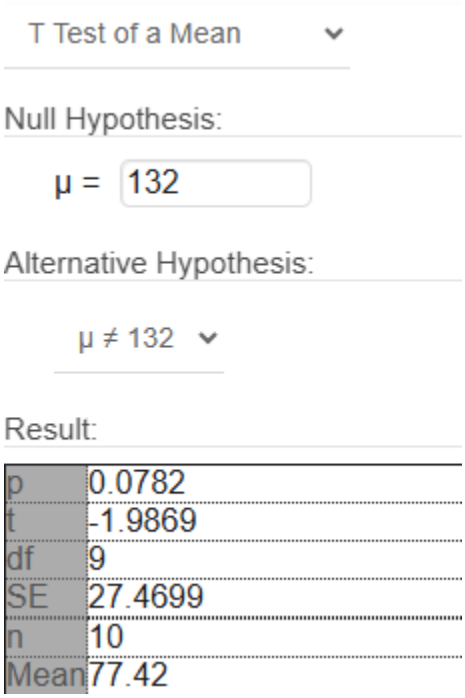
$1{,}146 \geq 10$ and $55{,}874 \geq 10$

Based on this information, I believe that the next steps will yield valid results.

Next, I am going to use GeoGebra to calculate a 95% confidence interval.

n	10
Mean	77.42
σ	82.4097
s	86.8674
Σx	774.2
Σx^2	127852.12
Min	4.8
Q1	9.2
Median	52.5
Q3	72
Max	244

Now, I want to see if the incidence rate of TB in the world is different than the expected 132 people per 100,000. For this I will be performing a T test of a means in GeoGebra to evaluate if $H_0: \mu = 0.00132$ and $H_A: \mu \neq 0.00132$. For this test, I will use $\alpha = 0.001$.



In the image above, you can see the p-value of 0.0782 and the test statistic of -1.9869. Because P>A, we will fail to reject the null hypothesis because there is not enough evidence to support the claim.

Travis Hornbuckle

Part IV – Tuberculosis Project

Introduction to Statistics (MATH1040)

Member of WHO	Success	Failure	Total
South Sudan	82	18	100
Botswana	76	24	100
Colombia	74	26	100
Guyana	70	30	100
Singapore	73	27	100
Australia	86	14	100
Azerbaijan	82	18	100
Czech Republic	63	37	100
Egypt	87	13	100
Libya	60	40	100
TOTAL	**753**	247	**1000**

B-1) For this portion, we want to find the probability that a randomly selected case is from Singapore **or** Australia. For this we would divide the total number of cases of both countries – in this example they're both 100 cases each, so 100*2 =200, then divide that by the total number of cases, which is 1000. This gives us a probability of 20% that a randomly selected case is from Singapore or Australia.

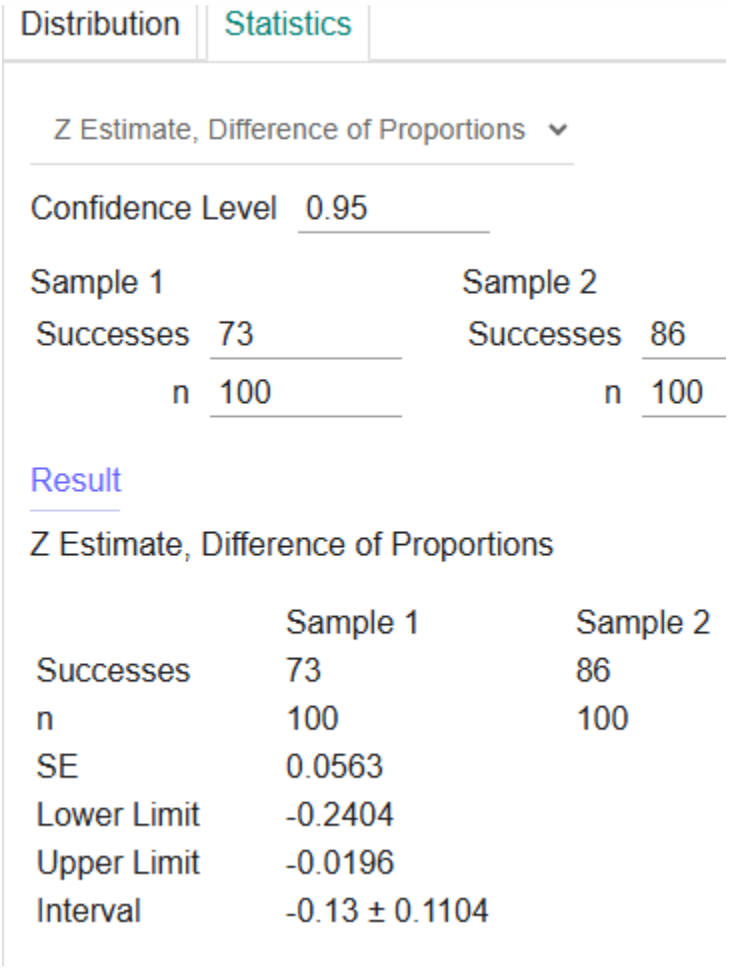
B-2) Now, we want to fin the probability that a case is either a failure or is from Singapore. To do this we would add all the failures, which is 247, then subtract the number of cases from Singapore – because we are looking for **OR**, not **AND**, which gives us a value of 220 cases. We then add the number of successes from Singapore, which is 73. Bringing us to a total of 293. That divided by the total number of cases gives us a probability of 29.3% that a randomly selected case is from Singapore **or** is a failure.

B-3) But what if we only wanted to see the proportion of cases that are from Singapore **AND** are failures? That would mean we take the 27 cases and divide that by the total number of cases, giving us a proportion of 2.7%.

B-4) What if we were only evaluating failures? What is the likelihood that a randomly selected case comes from Australia? To find that we would take the number of failures from Australia and divide that by the total number of failures, or 14/247, which gives us a proportion of 5.668%.

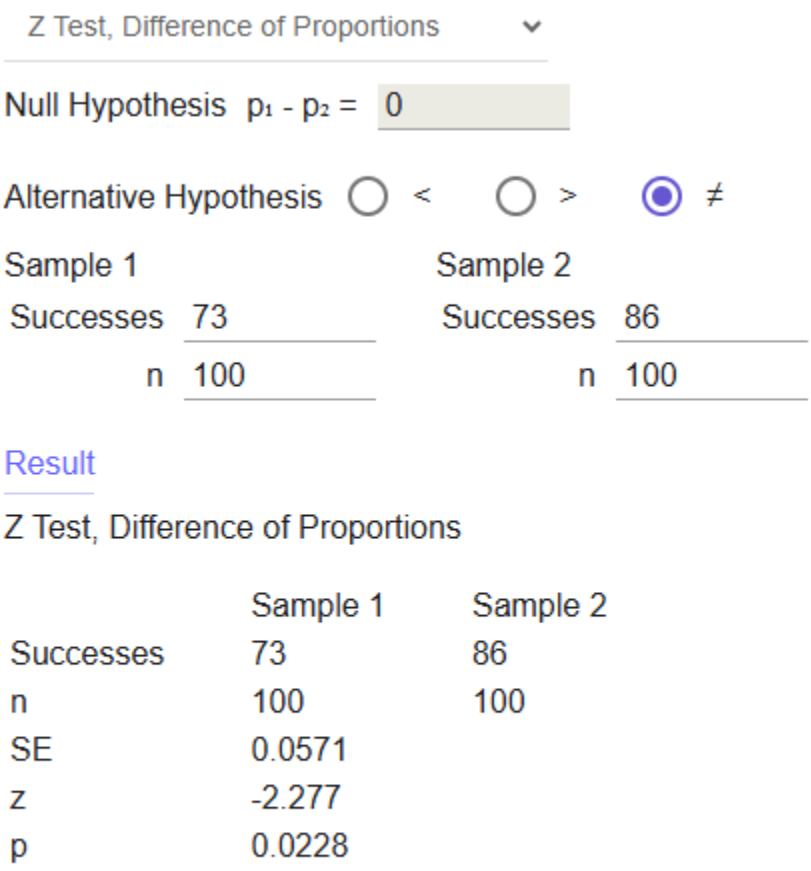
B-5) And if we wanted to find the odds of three randomly selected cases all coming from the Czech Republic? To find that, we would take the total number of cases coming from the Czech Republic, and divide it by the total number of cases – giving us the countries proportion, or 10% probability. We would then cube that to find the probability that it would happen 3 times, or $(0.1)^3$, which gives us a probability of 0.1% probability of all three cases coming from the Czech Republic.

C) To ensure that out information is accurate, we want to calculate a 95% confidence interval for Singapore and Australia. That would look like this when entering computations in GeoGebra as a Z Estimate, Difference of Proportions:



D) With these two countries, now we want to test them against our original hypothesis. Our null hypothesis is $H_0: p = 0.85$ – which is the baseline treatment rate for developed

countries according to WHO. Our alternative hypothesis is $H_A: p \neq 0.85$. For this test, I want to use an alpha of 0.05. In GeoGebra, my inputs look like this:



In the image above, we can see that the t-value is -2.277, and our p-value is 0.0228. With the alpha that I selected earlier we would compare those two and find that α > p, therefore our results are statistically significant. Based on this information we can conclude that there is sufficient evidence to support the hypothesis that $p \neq 0.085$, and therefore reject the null hypothesis – assuming all conditions are met.